DOCUMENT RESUME

ED 217 723 FL 013 013

AUTHOR Perkins, Kyle
TITLE An Analysis of the Robustness of Composition Scoring Schemes.
PUB DATE May 82
NOTE 16p.; Paper presented at the Annual Convention of Teachers of English to Speakers of Other Languages (16th, Honolulu, HI, May 1-6, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Diagnostic Tests; *English (Second Language); Higher Education; *Holistic Evaluation; Language Tests; Multiple Regression Analysis; *Testing; *Writing (Composition); *Writing Evaluation
IDENTIFIERS *Analytical Scoring

ABSTRACT

Various claims have been made about the efficiency of different composition scoring schemes. This paper reports the results of a systematic analysis of the robustness of holistic scoring, analytical scoring, and an obective diagnostic test, Davidson's "Test of the Ability to Subordinate." Compositions from an advanced English as a Second Language composition course were assigned holistic scores by two independent raters. The following variables were quantified for each paper: analytical scores, ratings on grammar, fluency, relevance, vocabulary, and mechanics. In addition, data from the "Test of Ability to Subordinate" were also collected. The data were submitted to a multiple regression analysis, the holistic scores being the dependent variable and all other scores being the predictor variables. The analytical variable accounted for 89% of the variance in the holistic scores. The only other variables to approach significance in accounting for variance in the dependent variable were ratings on vocabulary and fluency. It was not expected that a proficiency metric and a diagnostic metric would yield the same results. A further study is called for employing two different evaluator pools in order to confirm or disconfirm the findings of the present study. (Author/AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED217723

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Kyle Perkins
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AN ANALYSIS OF THE ROBUSTNESS OF COMPOSITION SCORING SCHEMES

Kyle Perkins

Southern Illinois University at Carbondale

Summary

Various claims have been made about the efficiency of different composition scoring schemes. This paper reports the results of a systematic analysis of the robustness of two different scoring schemes--holistic scoring and analytical scoring--and an objective diagnostic test--Davidson's (1978) Test of Ability to Subordinate.

Compositions from an advanced ESL composition course were assigned holistic scores by two independent raters. The following variables were quantified for each paper: analytical scores; ratings on grammar, fluency, relevance, vocabulary and mechanics. In addition, data from the Test of Ability to Subordinate were also collected.

The data were submitted to a multiple regression analysis, the holistic scores being the dependent variable and all other scores being the predictor variables. The analytical variable accounted for 89% of the variance in the holistic scores. The only other variables to approach significance in accounting for variance in the dependent variable were ratings on vocabulary and fluency.

FL013013

There are many procedures and instruments available for composition evaluation, but only three will be treated in this paper--holistic

scoring, analytical scoring, and an objective, diagnostic test, Test of Ability to Subordinate.

Holistic scoring involves one or more readers awarding a single grade based on the total impression of the composition as a whole text or discourse. My own personal bias is that holistic scoring has the highest construct validity of any composition evaluation scheme, when overall attained writing proficiency is the construct to be assessed. The bases for this position are the following: the whole text is valuated; the grader(s) can determine whether a thesis or issue has been clearly stated and developed or resolved; whether sufficient support and development have been provided for the reader; whether the writer has tried to accommodate the reader in terms of how much (little) the writer thinks the reader knows about the subject; whether there is grammatical and lexical cohesion and overall coherence; whether one paragraph is "hinged" smoothly to preceding and succeeding paragraphs through the accomplished use of transitional words, phrases, and structure vocabulary; whether the piece "makes sense."

Although I value holistic scoring highly for proficiency testing, I am aware that the method has its drawbacks and potential and realized weaknesses. As previously mentioned, with holistic scoring, the grader reads the composition, forms a general impression and assigns a mark to that composition based on some standard. That standard may be a model composition to which the reader has reference or some vague notion in his/her mind. Such evaluation can be highly subjective due to bias, fatigue, internal lack of consistency, previous knowledge of the student,

and/or shifting standards from one paper to the next.

Published research on holistic scoring in terms of reliability is a mixed bag. Braddock, Lloyd-Jones and Schoer (1963) and Follman and Anderson (1967) reported interrater reliability coefficients as high as .90. Kaczmarek (1980) found that subjective teacher judgments have substantial reliability and strongly correlated with independent readers and with objective scores.

Yet, other research indicates that professional persons, including English teachers, vary in their assessment of attained writing proficiency. The research of Remondino (1959), Diederich et al (1961) and Diederich (1974) is most noteworthy. Diederich et al (1961) reported a study in which sixty professionals were asked to grade 300 papers written by college freshmen from three different schools. The readers, who represented six occupational fields including college English teachers, social science teachers, natural science teachers, writers and educators, lawyers, and business executives, were asked to sort the 300 papers into nine groups.

Diederich (1974:5) noted that "out of the 300 essays graded, 101 received every grade from 1 to 9, 94 percent received either seven, eight, or nine different grades; and no essay received less than five different grades from fifty-three readers."

Diederich analyzed the reasons for the diversity of judgment and isolated the following predominant criteria: flavor and personality; organization and analysis; quality of ideas; usage, sentence structure, punctuation and spelling; wording and spelling.

Hirsch (1977), searching for reasons for lack of agreement among graders, suggested that different weights are attached to a few traits of writing and further noted that reliable, independent agreement in the scoring of writing samples is out of the question until widespread agreement is obtained about the qualities of good writing.

Objective measures such as words per T-unit, clauses per T-unit, etc. are so called and are said to be reliable, because independent evaluators will exhibit high correlations with each other when quantifying objective measures in writing work samples. Results from objective measures studies vary from researcher to researcher. For example, Arthur (1979), in a study of short-term changes in EFL composition skills, found the most significant changes in writing speed and in vocabulary size. No significant differences were noted in words per T-unit nor in error-free T-units. In the same volume, Kameen (1979) reported that 25 "good" and 25 "poor" writers differed in T-unit length, clause length and incidence of dynamic passive constructions. Perkins (1980) used ten objective measures of writing proficiency to evaluate ESL compositions. Only those objective measures which took the absence of errors into account discriminated among holistic evaluations of the compositions by experienced ESL teachers; and those measures were error-free T-units per composition, words per error-free T-unit, errors per T-unit, and total errors. Using very conservative statistical procedures, Perkins and Homburg (1979) found that total errors and errors per T-unit would discriminate between three different holistic evaluations. Flahive and Snow (1980), conducting a discriminant function analysis of several objective measures, reported that length of T-unit and clause per T-unit ratio were the only objective

measures studied by them which would discriminate among different proficiency levels of writing.

Only one diagnostic test was employed in this research--Test of Ability to Subordinate. According to Davidson's (1978) Teacher's Manual,

> The Test of Ability to Subordinate was designed primarily as a diagnostic instrument to measure the ability of non-native speakers of English to control certain structures of subordination in writing It was developed to fill the need for an easily scoreable, objective test that would measure students' ability to manipulate particular grammatical structures generally regarded as indicative of mature writing.... The primary diagnostic purpose of the TAS is to identify which of nine structures of subordination students can manipulate and which they cannot, but a careful analysis of responses will highlight difficulties, if any, with subject-verb agreement, plurals, tenses and articles. A teacher can use the information from the TAS to group students in terms of their strengths and weaknesses in subordination and to develop a writing program that will enable them to use these structures effectively as well as deal with other grammatical problems. (Davidson 1978:1).

Procedures

Subjects and Methodology

This research was conducted in the Department of Linguistics at Southern Illinois University. The purpose of the study was to determine which predictor variables would account for the greatest amount of variance in holistic scores. Holistic scores were chosen as the criterion, because I believe holistic scoring has the highest construct validity for assessing attained writing proficiency. Total analytical scores; ratings on relevance, fluency, vocabulary, grammar, and mechanics; an error count, words per T-unit and TAS objective scores were chosen as the predictor variables. The choices were made on the basis

of my previous pilot research and the published research of others.

The subjects were 102 foreign students enrolled in five sections of freshman composition for foreign students in the Department of Linguistics. Linguistics 101 is a substitute (for foreign students) for the English Department's freshman composition course. Seventeen native languages were represented in the subject pool: Laotian, Arabic, Farsi, Malay, Korean, Vietnamese, Thai, Chinese, Japanese, Efik, Ibo, Yoruba, Greek, Portuguese, Spanish, Norwegian and Gujarati. The instructors were graduate assistants.

The TAS was administered during the first and last weeks of the 15-week term as part of another research project. At the end of term, a final composition was assigned to all sections.

Before any grading was done, the graduate assistants were given training sessions on holistic and analytical scoring, before the compositions analyzed for this project were marked. Both the holistic and analytical scoring were conducted by the graduate assistants, my colleague Beverly H. Konneker and myself. The holistic grading and analytical grading were counterbalanced.

For the holistic grading we used a 5-point scale with the following descriptors:

5: excellent	The essay matches the top standard of native writers.
4: very good	The essay exhibits only a few minor problems

3: good — The essay communicates something of the writer's ideas, but with frequent interference; and/or there is clear communication, but with rather simple presentation.

2: poor — The essay exhibits serious problems; only half or less of the elements and ideas are coming through; the reader has to struggle.

1: bad — The essay is almost incomprehensible, difficult to process, and/or is excessively simple, containing only the bare elements.

The following evaluation form and descriptors were used for the analytical scoring:

	5	4	3	2	1
Relevance					
Fluency					
Vocabulary					
Grammar					
Mechanics					

(Maximum 25)

RELEVANCE means adherence to the topic, thesis development with supports, and overall unity.

FLUENCY refers to readability, organization, and coherence.

VOCABULARY refers to level of word choice, variety of vocabulary, appropriate usage, and idioms.

GRAMMAR refers to cohesive markers, the complexity of sentence structures, and sentence boundaries (e.g., no fragments or run-ons).

MECHANICS include minor grammar rules (e.g., subject-verb agreement), spelling and punctuation.

The holistic and analytical scoring were done independently and anonymously by each reader. The average of the two reader's scores for each paper was recorded for the holistic, total analytical and analytical subscores, i.e., ratings on relevance, fluency, vocabulary, grammar and mechanics. The interrater correlations ranged from .47 to .77 with a mean correlation of .58, yielding 34% true overlap between graders--$(.58)^2=.34$. I am not particularly proud of these coefficients, but I report them without any attenuation. Obviously, these coefficients are closer to those reported by Diederich than those reported by Braddock et al.

Analysis

To determine the variables that account for the most variance in the holistic scores, the data were submitted to the maximum R^2 Improvement technique from the SAS package. Maximum R^2 Improvement is one technique in a stepwise regression which is an attempt to search for the "best" model by bringing into the regression equation the independent variables one by one. R^2 is the test statistic generated by this technique, and it measures how much variation in the dependent variable that can be accounted for by the model. The operation of the

technique is explained below:

> This technique does not settle on a single model. Instead, it looks for the "best" one-variable model, the "best" two-variable model, and so forth. It finds first the one-variable model producing the highest R^2 statistic. Then another variable, the one which would yield the greatest increase in R^2 is added. Once this two-variable model is obtained, each of the variables in the model is compared to each variable not in the model. For each comparison, the procedure determines if removing the variable in the model and replacing it with the presently excluded variable would increase R^2. After all the possible comparisons have been made, the switch which produces the largest increase in R^2 is made. Comparisons are made again, and the process continues until the procedure finds that no switch could increase R^2. The two-variable model thus settled on is considered the "best" two-variable model the technique can find. The technique then addes a third variable to the model, according to the criteria used in adding the second variable. The comparing-and-switching model is discovered, and so forth (Barr, Goodnight, Sall and Helwig 1976:251-2).

The results of the regression analysis to determine the amount of variance in the holistic evaluations are presented in the following table.

REGRESSION ANALYSIS

HOLISTIC SCORES-DEPENDENT VARIABLE

Step	Variable Entered	R-Square	R-Square Improvement
1	Analytical Scores	.894	.894
2	Vocabulary	.931	.037
3	Errors	.947	.016
3	Vocabulary replaced by words/T-unit	.953	.006
4	Errors replaced by pre-TAS	.981	.028
5	Grammar	.983	.002
5	Pre-TAS replaced by Relevance	.986	.003
6	Grammar	.988	.002
7	Post-TAS	.989	.001
8	Pre-TAS	.989	.000
9	Vocabulary	.989	.000

Analytical scores is the "best" one-variable model accounting for 89% of the variance in the dependent variable. Analytical scores and vocabulary rating constitute the "best" two-variable model, together the two accounting for 93% of the variance in the dependent variable, with vocabulary accounting for 3.7% of the dependent variance. Errors replaced by pre-TAS (Step 4) accounts for an additional 2.8% of variance.

Since there are several variables in the equation, I use the following rule-of-thumb: if at least 2% of the variance in the dependent variable was not accounted for, the regression was not significant. Using this seat-of-the-pants criterion, one can see that only steps 1, 2, and 4 were significant meaning that the analytical scores, vocabulary ratings, and pre-TAS were accounting for significant amounts of variance in the dependent variable.

Discussion

It is rather obvious that all the graders were essentially looking for the same features in both the holistic scoring and the analytical scoring, and consequently this finding is a defect of the study. I say that a defect exists--both scoring procedures having high overlap--, because, traditionally holistic scoring has been highly touted for proficiency and placement purposes while analytical scoring has been recommended for diagnostic scoring; one would not expect to find two such different scoring methods measuring the same construct(s), whatever it (they) are. During the evaluation period the two scoring procedures were counterbalanced, done independently

and anonymously in an attempt to preclude any halo effect, contamination or biases from affecting the results. Were I to replicate this study, I would have two independent pools of evaluators, one to do the holistic scoring; the other to do the analytical scoring.

I have no explanation to offer for why the vocabulary rating accounted for a significant amount of variance unless the raters valued this trait more highly than other traits. In other previously-conducted research I found grammar always to be a robust predictor.

The fact that the pre-TAS accounts for a significant portion of variance in the dependent variable lends some credence to any claim that might be made for its predictive validity. The TAS comes from a school of thought that holds that enhancement of syntactic fluency will also be reflected in improvement in overall writing ability. To a certain degree that hypothesis was confirmed by step 4 of the regression equation.

In summary, the results of this study are perplexing. Evaluators using two different scoring procedures, in essence, assessed the same constructs. One would not expect a proficiency metric and a diagnostic metric to yield the same results. A further study employing two different evaluator pools would confirm or disconfirm the findings of the present study. Second, there is evidence that evaluators value various traits with differential weights, e.g., observe steps 2, 3, 5, and 6. Third, there is some indication that the TAS has predictive validity but little or no concurrent validity, e.g., observe steps 4 and 7. And finally, there is no indication that the objective

measure employed in this study has any concurrent or construct validity for attained writing proficiency, e.g., observe step 3. This finding is not at all surprising, because there is universal agreement that the purpose of writing is to communicate information, and objective measures such as words per T-unit have no necessary or sufficient connection with a writer's intent to communicate. To invoke testing jargon, I would add that objective measures may exhibit high coefficients of estimates of reliability, but they do not exhibit construct validity for oral proficiency testing in writing.

REFERENCES

Arthur, Bradford. 1979. Short-term changes in EFL composition skills. In Carlos A. Yorio, Kyle Perkins, and Jacquelyn Schachter (eds.). On TESOL '79: The Learner in Focus. Washington, D.C.: Teachers of English to Speakers of Other Languages, 330-342.

Barr, Anthony, James H. Goodnight, John P. Sall and Jane T. Helwig. 1976. A User's Guide to SAS 76. Raleigh, North Carolina: SAS Institute, Inc.

Braddock, R., R. Lloyd-Jones and L. Schoer. 1963. Research in Written Composition. Urbana, Illinois: National Council of Teachers of English.

Davidson, David M. 1978. Test of Ability to Subordinate. New York: Language Innovations, Inc.

Davidson, David M. 1978. Teacher's Manual for Test of Ability to Subordinate. New York: Language Innovations, Inc.

Diederich, P., J. W. French, and S. Carlton. 1961. Factors in judgments of writing ability. Educational Testing Service Research Bulletin. Princeton, New Jersey: Educational Testing Service.

Flahive, Douglas E. and Becky Gerlach Snow. 1980. Measures of syntactic complexity in evaluating ESL compositions. In John W. Oller, Jr. and Kyle Perkins (eds.). Research in Language Testing. Rowley, Mass.: Newbury House Publishers, Inc., 171-176.

Follman, J. C. and J. A. Anderson. 1967. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English 1:190-200.

Hirsch, E. D., Jr. 1977. The Philosophy of Composition. Chicago: The University of Chicago Press.

Kaczmarek, Celeste M. 1980. Scoring and rating essay tasks. In John W. Oller, Jr. and Kyle Perkins (eds.). Research in Language Testing. Rowley, Mass.: Newbury House Publishers, Inc., 151-159.

Kameen, Patrick T. 1979. Syntactic skill and ESL writing quality. In Carlos A. Yorio, Kyle Perkins and Jacquelyn Schachter (eds.). On TESOL '79: The Learner in Focus. Washington D.C.: Teachers of English to Speakers of Other Languages, 343-350.

Perkins, Kyle. 1980 using objective measures of attained writing proficiency to discriminate among holistic evaluations. TESOL Quarterly 14 1:61-69.

Perkins, Kyle and Taco J. Homburg. 1980. Three different statistical analyses of objective measures of attained writing proficiency. In Raymond O. Silverstein (ed.), Occasional Papers on Linguistics No. 6. Proceedings of the Third International Conference on Frontiers in Language Proficiency and Dominance Testing. Carbondale, Illinois: Department of Linguistics, Southern Illinois University, 306-317.

Remondino, C. 1959. A factorial analysis of the evaluation of scholastic compositions in the mother tongue. British Journal of Educational Psychology 30:242-251.